A NOTE ON COMPUTER SYSTEM DATA GATHERING

Jack P.C. Kleijnen
Katholieke Hogeschool
Tilburg, Netherlands

Recently Orchard (1977) proposed a statistical technique for data collection in computer systems. A main idea was the use of random sampling, as opposed to traditional fixed periodic sampling. He further proceeded to derive confidence intervals for the resulting estimator. He also proposed the use of binary (Boolean) variables, e.g., $q_{it} = 1$ (or 0) if at sampling time t the ith "slot" of a queue is occupied (or empty respectively).

Unfortunately, as I understand the author's exposé, the derived confidence intervals depend on the assumption of independent observations. This assumption, however, is violated in dynamic systems such as computer systems (or their corresponding simulation models).

For instance, suppose that at $t=t_1$ the system is heavily loaded, so that $q_{it_1} = 1$ for all i-values. Sample the next sampling moment, say $t=t_2$ (Orchard proposed to make t uniformly distributed; see pp. 33-34) Suppose t_2 turns out to be slightly larger than t_1. Then the probability that, say, $q_{1t2} = 1$ is higher than it would have been if the system were lightly loaded at $t=t_1$. In other words, the observations on q_{it} are serially correlated! More generally, if q_t is a time series, then in whatever order we observe all or some of these q_t, we are confronted with serial correlation.

In Kleijnen (1975, 454-468) three alternatives are discussed for tackling the autocorrelation problem:

(1) Estimate the serial correlation coefficients.

(2) Take the observations "sufficiently" far apart, so that the dependence may be ignored.

(3) Make observations during "epochs" which are independent because of the "renewal" property of certain stochastic systems.

Besides the variability of the estimator one should consider the bias of the estimator. Observing a stochastic process at fixed or uniformly distributed points of time may create bias, if the process is not Markovian (Poisson arrivals and services in a queuing system). This can be seen intuitively in case the process

(continued on page 62)

DRUM-2

Correlation Coefficient = 0.92372

STD. Error = 1.4405

Prediction Equation

$$Z = -3.0591 + 30.8241X^3 + 1.3222.P + 3.2817\ XM - 11.5977.X + 0.055026\ X.M.P - 0.059479P^2$$

DRUM-3

Correlation Coefficient = 0.75131

STD. Error = 3.36743

Prediction Equation

$$Z = 4.8829 + 44.3342X^3 + 0.00124\ X.M.C + 0.04877\ X.C + 0.023718C - 0.003622\ M^3 + 0.001363M.C$$

Legend: Z = Throughput (no. of jobs/unit time)
M = Multiprogramming Level
P = No. of Pages of Memory (1K. Page Size)
S = Paging Speed (in sec)
X = Job Mix (Percent)

REFERENCES

1. Abate, J., H. Dubner, and S. B. Weinberg, "Queuing Analysis of the IBM 2314 Disk Storage Facility", JACM, 15, 4 (1968).
2. Coffman, E. A. and T. A. Ryan, "A Study of Storage Partitioning Using Mathematical Model of Locality," CACM, Vol. 15, No. 3, 1972.
3. Coffman, E. G. and L. C. Varian, "Further Experimental Data on the Behavior of Programs in a Paging Environment", CACM, Vol. 11, 5, 1968.
4. Denning, P. J., "The Working Set Model for Program Behavior", CACM, 11, 5, 1968.
5. Denning, P. J., J. E. Savage and J. R. Spirn, "Models of Locality in Programs Behavior," TR-107, Dept. of Electrical Engineering, Princeton University (1972).
6. Fine, G. H., C. W. Jackson and P. V. McIssac, "Dynamic Program Behavior Under Paging", Proc. Natl. ACM, 21st, (1966).
7. Kuck, D. J. and D. H. Laurie, "The Use and Performance of Memory Hierarchies: A Survey", Software Engineering, Vol. 1, Julius Ton, Ed., Academic Press (1970).
8. Rodriguez-Rosell, J., "Experimental Data on How Programs Behavior Affects Choice of Scheduling Parameters", Proc. 3rd ACM Symp. on o/s Princ. (1971).
9. Seaman, P. H., R. A. Laird and T. L. Wilson, "On Teleprocessing System Design. Pt. IV.-- Analysis of Auxiliary Storage Activity", IBM System J., Vol. 5, No. 3, 1966.
10. Shedler, G. S. and C. Tung, "Locality in Page Reference Strings", SIAM J. Comput. 1, 3 (1972).
11. Smith, J. C., "Multiprogramming Under Page on Demand Strategy", CACM, 10 (1967).
12. Teorey, T. J. and T. B. Pinkerton, "A Comparitive Analysis of Disk Scheduling Policies", Proc. 3rd Symp. O/S Princ. (Oct. 1971).

A NOTE ON COMPUTER SYSTEM....

(continued from page 56)

shows periodic behavior. To obtain unbiased measurements the"interarrival times" between sampling points should be exponentially distributed: Poisson measurement process.

Another issue that deserves mentioning is sequential sampling. For instance, since σ in Orchard's eq. (3) is unknown, one may start sampling, compute an estimate s^2, substitute this estimate into eq. (3) continue sampling, update s^2, etc. This more efficient approach(and several variants) is discussed at length in Kleijnen (1975, pp. 479-506). Note that sequential sampling also applies to binary variables.

Stratified sampling briefly discussed by Orchard, is further analyzed in Kleijnen (1975, pp. 110-133). However, other variance reduction techniques may be more attractive, e.g. control variates; see Kleijnen (1975, p.p. 105-285).

References:

1. Kleijnen,J.P.C., STATISTICAL TECHNIQUES IN SIMULATION.(In two volumes) Marcel Dekker Inc., New York, 1974/1975.
2. Orchard, R.A., A new methodology for computer system data gathering. PERFORMANCE EVALUATION REVIEW, 6, no.4, Fall 1977, pp. 27-41.